W624–W627 *Nucleic Acids Research, 2004, Vol. 32, Web Server issue*
DOI: 10.1093/nar/gkh486

CSTminer: a web tool for the identification of coding and noncoding conserved sequence tags through cross-species genome comparison

Tiziana Castrignanò, Alessandro Canali, Giorgio Grillo[1], Sabino Liuni[1], Flavio Mignone[2] and Graziano Pesole[2,*]

Consorzio Interuniversitario per le Applicazioni di Supercalcolo per Università e Ricerca, CASPUR, Rome, Italy, [1]Istituto Tecnologie Biomediche, Sezione di Bioinformatica e Genomica, Consiglio Nazionale delle Ricerche, Bari, Italy and [2]University of Milan, Dipartimento di Scienze Biomolecolari e Biotecnologie, via Celoria 26, Milan 20133, Italy

Received February 14, 2004; Revised April 15, 2004; Accepted May 5, 2004

ABSTRACT

The identification and characterization of genome tracts that are highly conserved across species during evolution may contribute significantly to the functional annotation of whole-genome sequences. Indeed, such sequences are likely to correspond to known or unknown coding exons or regulatory motifs. Here, we present a web server implementing a previously developed algorithm that, by comparing user-submitted genome sequences, is able to identify statistically significant conserved blocks and assess their coding or noncoding nature through the measure of a coding potential score. The web tool, available at http://www.caspur.it/CSTminer/, is dynamically interconnected with the Ensembl genome resources and produces a graphical output showing a map of detected conserved sequences and annotated gene features.

INTRODUCTION

The identification of highly conserved noncoding sequences by pairwise sequence comparison of homologous genome loci may contribute greatly to the identification of a multitude of regulatory elements modulating extent, chronology and specific location of gene expression (1–3). A nucleotide similarity significantly above the background level—related to the divergence time between the organisms under investigation—indicates a selective constraint during evolution to preserve the functional role encoded in that specific sequence element. Such selective pressures can operate both on coding sequences as a consequence of functional constraints on the encoded proteins and on noncoding sequences involved in regulatory activity. Thus, we are faced with the problem of discriminating between conserved coding and noncoding sequences.

Indeed, although the complete genomes of several eukaryotic organisms, including humans, have been sequenced, we are not yet able to determine their complete gene inventory, i.e. the exhaustive delineation of coding regions, including alternative initiation, termination and splicing patterns of the corresponding transcripts.

In this context, a software tool to identify and assess the coding or noncoding nature of conserved sequence tags (CSTs) through cross-species genome comparisons would contribute to genome annotation through the identification of novel genes or gene expression isoforms. Such a tool would also facilitate the identification of transcribed or untranscribed noncoding regions likely to function as *cis*-regulatory elements.

We previously devised a simple method that is based on the quantification of a specific evolutionary pattern which is typical of coding regions, where synonymous substitutions outnumber non-synonymous ones at the nucleotide level and conservative changes outnumber non-conservative ones at the protein level. Thus, we generate a coding potential score (CPS) for each CST identified in a pairwise genome comparison (4). CSTs with a low coding potential have been shown to represent reliable candidates for noncoding regulatory motifs.

Here, we present a web tool which implements this method and which provides, through gene name or chromosome coordinates, direct access to Ensembl genomes (5) as well as the possibility for the user to submit data directly. A graphical output of detected CSTs also shows, where available, known gene features (mRNAs, exons, etc.) of the genomic region under investigation.

[*] To whom correspondence should be addressed. Tel: +39 02 50314915; Fax: +39 02 50314912; Email: graziano.pesole@unimi.it

SYSTEM AND METHODS

The web tool, accessible at http://www.caspur.it/CSTminer/, is implemented on a Linux server (RedHat 8.0) running apache web server version 2.0 (http://www.apache.org/). PHP scripts (http://www.php.net/) are used to provide access to the Ensembl database (http://www.ensembl.org/), to interact with the CSTminer application over the Internet and to plot dynamic web results. The dynamic plot is drawn using the GD libraries (version 1.6). A queue system is administered by a Perl script querying a MySQL database (version 3.23).

The web tool essentially implements the algorithm described in (4) with default parameters, i.e. word size of 7, maximum E-value of 10^{-5} for BLAST analysis and minimum CST length of 30 nt. In the present version of the algorithm, instead of the standard Blosum80 matrix previously adopted, we use a rescaled Blosum80 matrix with all amino acid interconversion scores made positive by the addition of a constant value (i.e. +8 in the case of Blosum80). In order to take into account the possibility of multiple hits, the CPS is now calculated using a slightly modified formula that uses a log transformation and replaces the ratio of observed synonymous/non-synonymous changes [see equation in (4)] with the K_s/K_a ratio calculated according to Nei and Gojobori (6):

$$\mathrm{CPS}(f) = \ln\left[\left(\frac{100L}{3}\right)\left(\frac{K_s}{K_a}\right)\sum_{i=1}^{N_f} AA\_sim_i\right].$$

The computed synonymous and non-synonymous changes are increased by one to avoid null denominators.

CSTminer also allows the display of the highest scoring triplet window (default minimum length of 60 nt) by scanning each detected CST. This approach facilitates the detection of potential coding regions located in longer CSTs which might contain both coding and noncoding tracts (through the presence of untranslated mRNA or intronic regions).

In the submission form the user submits two sequences to run jobs. This can be done in four different ways: (i) pasting the sequences (in FASTA format), (ii) uploading a text file containing one query sequence and one target sequence, (iii) submitting the Ensembl gene ID and selecting the corresponding organism and (iv) selecting the organism and choosing the chromosomal range. In the last two cases the web tool performs a query to the Ensembl database to extract the sequence of the gene (including surrounding upstream and downstream sequences) or sequences contained in the specified chromosomal range. The option for masking annotated repeats is also available in these cases.

After the job is submitted, a popup window opens and instructs the user to wait until the job is finished. On completion, a link guides the user to the dynamic web plot visualizing the CSTminer output. The plot shows the CSTs recovered in both orientations and highlights known gene features (obtained by performing a query to the Ensembl database related to the input sequences). A filter can be used to plot only the CSTs in a particular range of CPS values. When the cursor is placed over a CST a tooltip appears showing relevant summary information (CPS value, start, end); clicking on the CST opens a popup window showing all information (e.g. the CST sequence). When the cursor is moved over known features a tooltip shows the Ensembl name, its chromosome coordinates with respect to the query sequence and the hypertext link to the corresponding Ensembl gene.

APPLICATION

To test the effectiveness of the CSTminer method and to set up suitable cutoff values for the CPS, we analyzed two sets of sequences. First, we analyzed five RANDOM datasets, each one containing 10 000 pairs of random sequences 1000 nt long with the same dinucleotide composition as human 5′-UTR (Untranslated region) sequences. The sequence identity between pairs of sequences increased from 50–60% in the first dataset to 90–100% in the last dataset. Second, we analyzed a CDS dataset containing the coding regions of 1880 homologous pairs of human and mouse mRNAs (7).

To assess how well the artificial RANDOM dataset represented natural noncoding sequences we also analyzed a real noncoding dataset of 157 eukaryotic 18S rRNA sequences. We obtained 4844 rRNA CSTs with a CPS distribution (data not shown) and average value (CPS = 6.01) not dissimilar to that observed for the RANDOM dataset (average CPS = 6.09). We then used random sequences as the noncoding training set, as the greater number of sequences offered more statistically significant values.

A total of 34 100 and 2496 CSTs were detected for the RANDOM and CDS datasets, respectively, and the distribution of the corresponding CPSs is shown in Figure 1. The two distributions show only a limited overlap with the average CPS of coding CSTs, which at 8.51 is remarkably higher than that observed for noncoding CSTs. Of all the coding CSTs <1% showed a CPS <6.74, whereas <1% of the total noncoding CSTs showed a CPS >7.71. We arbitrarily classify CSTs into three classes as follows: CPS < 6.74, noncoding CST; $6.74 \leqslant \mathrm{CPS} \leqslant 7.71$, possibly coding CST; CPS > 7.71, coding CST. With reference to the assessment of the coding character, the estimated false positive rate is 9.85% with CPS > 6.74 and 1.00% with CPS > 7.71.

Figure 2 shows a snapshot of the result obtained by comparing the ferroportin genes of human and mouse. Colored bars represent detected CSTs, with different colors assigned to the three classes defined above; 7 out of 15 CSTs with high CPS are classified as coding, and 8 out of 15 with low CPS as noncoding. Indeed, all the former actually correspond to coding exons. Interestingly, the noncoding CST located in the 5′-UTR corresponds to a known functional regulatory motif involved in the iron-mediated post-transcriptional regulation of gene expression (8).

DISCUSSION

The completion of numerous eukaryotic and prokaryotic genome projects provides researchers with an extraordinary amount of valuable data. However, this information is not immediately fully interpretable using current experimental and bioinformatics approaches. In this context, comparative genome studies may be very valuable as the monodimensional information encoded in the genetic material is much easier to interpret if considered in its appropriate evolutionary framework. This can be achieved through comparison with homologous sequences in the same or other genomes. In

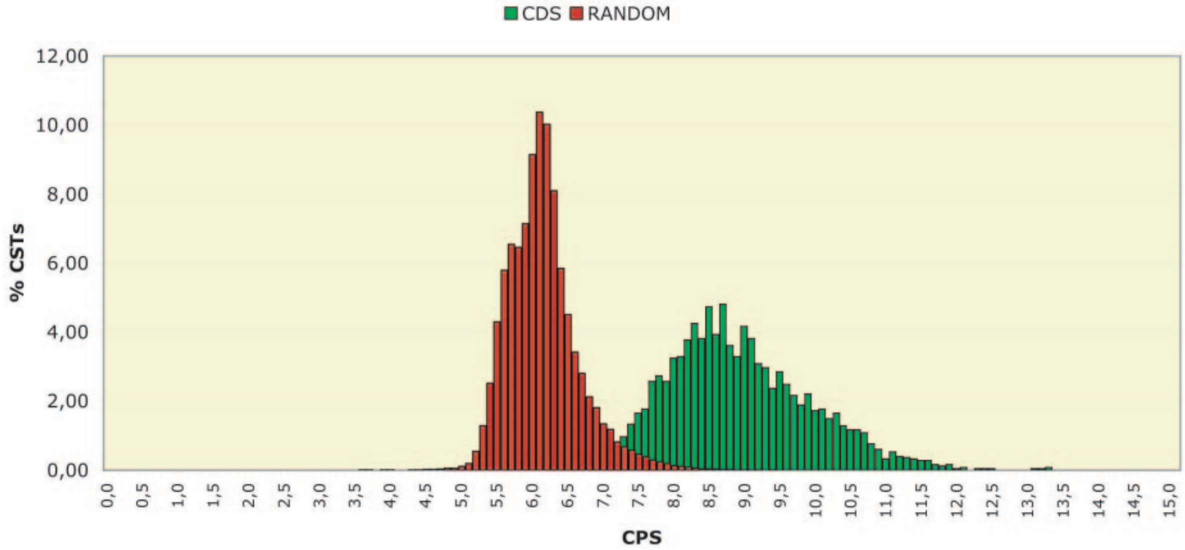
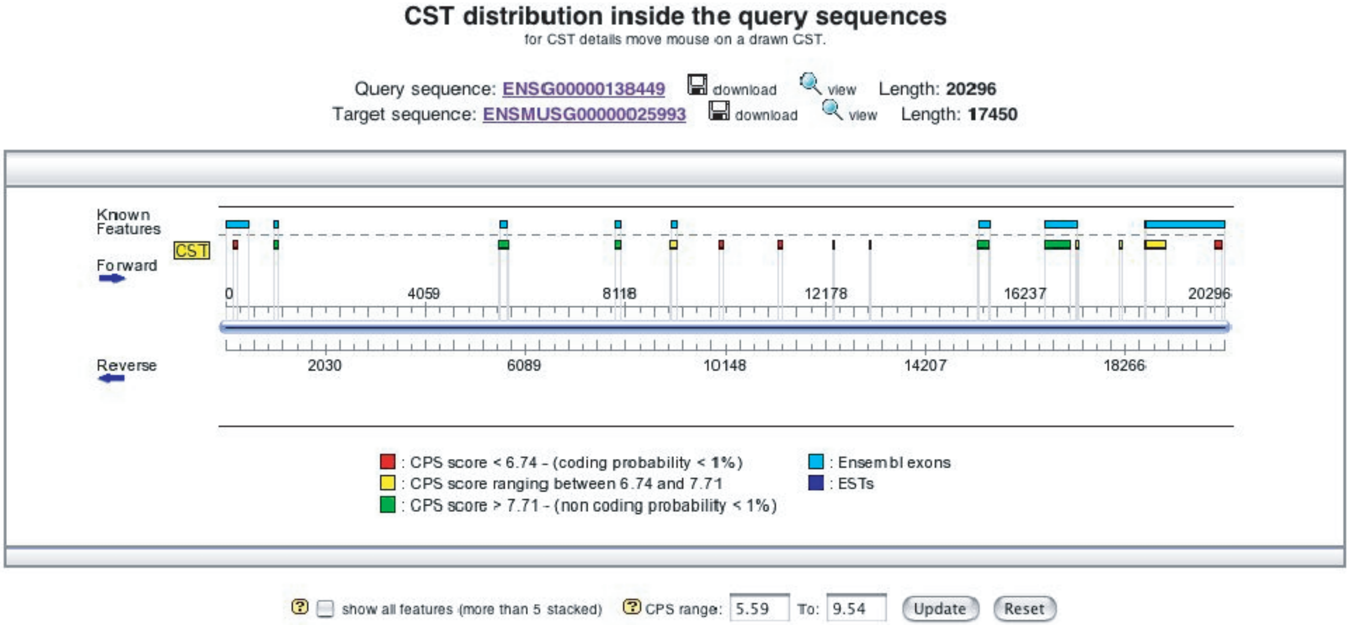


Figure 1. Distribution of CPS values for the RANDOM and the CDS datasets.



CST distribution inside the query sequences

for CST details move mouse on a drawn CST.

Query sequence: ENSG00000138449 download view Length: **20296**

Target sequence: ENSMUSG00000025993 download view Length: **17450**

🟥 : CPS score < 6.74 - (coding probability < 1%)

🟨 : CPS score ranging between 6.74 and 7.71

🟩 : CPS score > 7.71 - (non coding probability < 1%)

🟦 : Ensembl exons

🟦 : ESTs

? □ show all features (more than 5 stacked) ? CPS range: 5.59 To: 9.54 Update Reset

Figure 2. CSTminer output snapshot obtained by analyzing the ferroportin gene of human (Ensembl ID: ENSG00000138449) and mouse (Ensembl ID: ENSMUSG00000025993). The blue bars represent Ensembl exons annotated in the corresponding genomic region. When the cursor is moved over the exons, a tooltip gives some basic information (exon Ensembl ID, start, end), while an active link opens the Ensembl entry corresponding to the gene under investigation. Tooltips have also been implemented on the CST bars to provide a source of quick information about the CPS value and the relative position of the CST in the query sequence. Clicking the bar opens a new window showing more detailed information, including the CST sequence and its relative position in the target sequence. Clicking on the CST box opens a new window showing a summary table of all CSTs identified. The option is also given to view or download the query and target input sequences by clicking on the corresponding icons.

particular, selective pressures shape the evolution of DNA sequences as a function of the specific genetic information they encode. Thus, sequence regions coding for proteins undergo a strongly biased pattern of nucleotide changes, where synonymous substitutions significantly outnumber non-synonymous ones, and these latter more frequently result in conservative amino acid replacements. Analogously, we may expect specific covariation patterns in the evolution of DNA sequences coding for noncoding RNAs whose conserved secondary structure is under selective constraint (9,10).

Thus, the quality of genomic annotation provided by comparative sequence analysis may be improved greatly by taking the evolutionary dynamics of the compared sequences into proper account. Comparing genomic sequences of two or more genomes may contribute greatly to the identification of novel regulatory elements, novel genes or novel exons of alternatively spliced coding or noncoding genes, as these probably correspond to conserved sequence regions.

Local similarity blocks detected in the alignment of two genomic sequences are defined as CSTs if their length and identity are above an arbitrarily chosen cutoff, typically 100 nt and 70%, respectively (11). CSTminer, however, adopts a probabilistic cutoff, provided by the BLAST *E*-value. Thus, shorter CSTs with very high similarity, probably encoding relevant genetic information, are also detected. Once all of the CSTs of the genome tracts under investigation have been detected and mapped, the problem of their characterization (their assignment to any of the aforementioned functional classes) remains. To this end, we previously developed a simple method to provide a quantitative measure of the coding potential (CPS value) that proved to effectively discriminate between coding and noncoding CSTs with a low false positive rate (4).

The web server we present here implements an improved version of the method previously described (4) and allows researchers to investigate user-submitted genomic sequences as well as genes or genome regions automatically extracted from the Ensembl database (5) through a simple graphical interface. The functional annotation of detected CSTs can be easily inferred from a graphical output showing their location with respect to known features and sequence elements. The collection and characterization of putative noncoding CSTs associated with groups of genes whose products share properties such as cellular localization, Gene Ontology, functional domain and expression pattern may provide valuable insights for the detection of regulatory motifs and for the design of experimental validation procedures.

ACKNOWLEDGEMENTS

We thank Cecilia Saccone and David Horner for valuable comments on the manuscript. This work was supported by FIRB project Bioinformatica per la Genomica e la Proteomica, Cluster C03 ‘Studio di geni di interesse biomedico e agroalimentare’ (Ministero dell’Istruzione e Ricerca Scientifica, Italy) and by Telethon.

REFERENCES

1. Nobrega,M.A. and Pennacchio,L.A. (2004) Comparative genomic analysis as a tool for biological discovery. *J. Physiol.*, **554**, 31–39.
2. Pennacchio,L.A. and Rubin,E.M. (2001) Genomic strategies to identify mammalian regulatory sequences. *Nature Rev. Genet.*, **2**, 100–109.
3. Dermitzakis,E.T., Reymond,A., Lyle,R., Scamuffa,N., Ucla,C., Deutsch,S., Stevenson,B.J., Flegel,V., Bucher,P., Jongeneel,C.V. *et al.* (2002) Numerous potentially functional but non-genic conserved sequences on human chromosome 21. *Nature*, **420**, 578–582.
4. Mignone,F., Grillo,G., Liuni,S. and Pesole,G. (2003) Computational identification of protein coding potential of conserved sequence tags through cross-species evolutionary analysis. *Nucleic Acids Res.*, **31**, 4639–4645.
5. Birney,E., Andrews,D., Bevan,P., Caccamo,M., Cameron,G., Chen,Y., Clarke,L., Coates,G., Cox,T., Cuff,J. *et al.* (2004) Ensembl 2004. *Nucleic Acids Res.*, **32**(Database issue), D468–D470.
6. Nei,M. and Gojobori,T. (1986) Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol. Biol. Evol.*, **3**, 418–426.
7. Makalowski,W. and Boguski,M.S. (1998) Evolutionary parameters of the transcribed mammalian genome: an analysis of 2,820 orthologous rodent and human sequences. *Proc. Natl Acad. Sci. USA*, **95**, 9407–9412.
8. Lymboussaki,A., Pignatti,E., Montosi,G., Garuti,C., Haile,D.J. and Pietrangelo,A. (2003) The role of the iron responsive element in the ferroportin1/IREG1/MTP1 gene expression. *J. Hepatol.*, **39**, 710–715.
9. Rivas,E. and Eddy,S.R. (2001) Noncoding RNA gene detection using comparative sequence analysis. *BMC Bioinformatics*, **2**, 8.
10. Parsch,J., Braverman,J.M. and Stephan,W. (2000) Comparative sequence analysis and patterns of covariation in RNA secondary structures. *Genetics*, **154**, 909–921.
11. Pollard,D.A., Bergman,C.M., Stoye,J., Celniker,S.E. and Eisen,M.B. (2004) Benchmarking tools for the alignment of functional noncoding DNA. *BMC Bioinformatics*, **5**, 6.